

# Leveraging Infrared Spectroscopy for Automated Structure Elucidation

Marvin Alberts<sup>1</sup>, Teodoro Laino<sup>1,2</sup>, and Alain C. Vaucher<sup>1,2</sup>

<sup>1</sup>*IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

<sup>2</sup>*National Center for Competence in Research-Catalysis (NCCR-Catalysis), Zurich, Switzerland*

## Abstract

The application of machine learning models in chemistry has made remarkable strides in recent years. While the field of analytical chemistry has also received considerable interest from machine learning practitioners, very few models have been adopted into everyday use. Among the analytical instruments available to chemists, Infrared (IR) spectroscopy is one of the cheapest, easiest and most accessible. So far the use of IR has been limited to the identification of a select few functional groups with well-known vibrational frequencies with the interpretation of most peaks lying outside of human capabilities. We present a novel machine learning model that enables chemists to leverage the complete information contained within an IR spectrum to directly predict the molecular structure. To achieve this, we developed a transformer model trained on IR spectra that predicts the molecular structure as a SMILES string. To cover a vast portion of chemical space, we generated a training set of 634,585 simulated IR spectra using molecular dynamics. Our approach achieved a top-1 accuracy of 45.33% and a top-10 accuracy of 78.5% on a test set sampled from PubChem with a heavy atom count ranging from 6 to 13. The model is useful also in cases where an incorrect structure is predicted, as it is capable of predicting the correct scaffold in 77.01% of cases as the top-1 prediction and in 91.54% in the top-10 predictions. In addition, the model outperforms other models solely trained to predict the functional group from the IR spectrum.

## 1. Introduction

Infrared (IR) spectroscopy has been widely used in chemistry since the early 1900s when it was demonstrated that functional groups map to specific peaks in the spectrum [1].

The affordability of IR spectrometers have made them a staple in chemical laboratories, giving chemists a quick and easy way to identify which functional groups are present in a sample [2]. In forensics, pharmaceuticals, and food science, IR spectra are often used to identify chemical compounds by comparison to a database [3–6].

Despite its potential for automated structure elucidation, the need for an exhaustive database to match a spectrum remains a critical limitation in the effective utilization of IR spectra. This is exacerbated by the complexity and overlapping nature of spectral features that can impede accurate interpretation, particularly for human analysts. While identifying specific functional groups such as the carbonyl peak around  $1700\text{ cm}^{-1}$  is straightforward, decoding the fingerprint region ( $400\text{--}1500\text{ cm}^{-1}$ ) is a much more daunting task [2]. Consequently, chemists have traditionally utilised only a minimal amount of the information present in IR spectra to detect the presence of a handful of functional groups, leaving a significant portion of the spectrum’s potential for structure determination untapped.

However, the increased availability of more advanced methods such as NMR or LC-MS has reduced the importance of IR spectroscopy as a structure elucidation tool in research chemistry [7]. While NMR and LC-MS easily surpass IR spectroscopy in their structure elucidation power, they also have their limitations. NMR spectroscopy requires deuterated solvents, expensive instruments and measurement times ranging from 10 minutes to hours [8]. Similarly, LC-MS relies on expensive high-purity solvents, extensive method development, the sample is destroyed in the process while also requiring a database matching system for structure elucidation [9, 10]. IR spectroscopy, on the other hand, is quick, cheap, non-destructive, and easy to use.

With the rise of computing power, a wave of new statistical methods (i.e. machine and deep learning) have allowed tackling previously challenging problems such as image classification or language modeling [11, 12]. Machine learning and especially language modeling has shown great promise in chemistry. Applications of such models range from predicting retrosynthetic routes, over designing novel drug candidates to aiding in the automation of experiments [13–15]. In the field of IR spectroscopy, machine learning has also advanced the processing of spectra. Convolutional neural networks (CNNs) have achieved state-of-the-art performance on predicting functional groups from IR spectra [16, 17]. Other types of machine learning models, such as support vector machines (SVMs), random forest models, and multilayer perceptrons (MLPs), have also been employed to predict functional groups from IR spectra [18–20].

Despite the pressing need for rapid and accurate structure elucidation methods in chemistry, direct prediction of the complete chemical structure from IR spectra has yet to be accomplished, even with the recent advances in machine learning. Unlocking this capability will enable chemists to fully utilise the wealth of information present in IR spectra and breathe new life into the use of IR spectroscopy in analytical chemistry. Additionally,

such a fast and cost-effective elucidation tool will have broad applicability across various fields, ranging from research chemistry over metabolomics to forensics.

Here, we present the first work using a machine learning model for full structure elucidation from the IR spectrum. We use a transformer model trained on both the molecular formula and the IR spectrum to directly predict the molecular structure as Simplified molecular-input line-entry system (SMILES) [21]. IR spectra are simulated using molecular dynamics and the PCFF force field [22]. We evaluate the model’s ability to predict the correct molecule, scaffold and functional groups from the IR spectrum. Our model achieves a 45.33% top-1 and 78.5% top-10 accuracy while predicting the correct structure, 77.01% top-1 and 91.54% accuracy to predict the scaffold and an average F1 score of 0.961 when predicting 21 functional groups.

## 2. Results and Discussion

### 2.1. Model

Our model adopts a sequence-to-sequence transformer architecture. The input comprises the molecular formula and the IR spectrum, while the output is represented as SMILES and denotes the molecular structure (see Figure 1). Transformers have a proven performance in generating molecules as SMILES [23,24], and have recently been shown to excel at processing numbers as well [25,26]. More details on the architecture are given in the Methods section 4.

### 2.2. Model optimisation

We trained a total of 20 models, employing different data augmentation techniques and varying the inclusion of the chemical formula, token length, and section of the IR spectrum. For each spectrum in the test set, we generated ten ranked predictions and calculated the accuracy of each model by comparing the predicted structures to the target structure. Specifically, we report the top-1, top-5, and top-10 accuracy metrics, which indicate the percentage of cases where the predicted structure matches the target structure within the first, first five, and first ten predictions, respectively. Two molecules are defined as matching if their canonical SMILES strings match exactly. Table 1 shows the results of the trained models.

In the following, we delve deeper into the methodological choices adopted for data preparation and their respective effects.

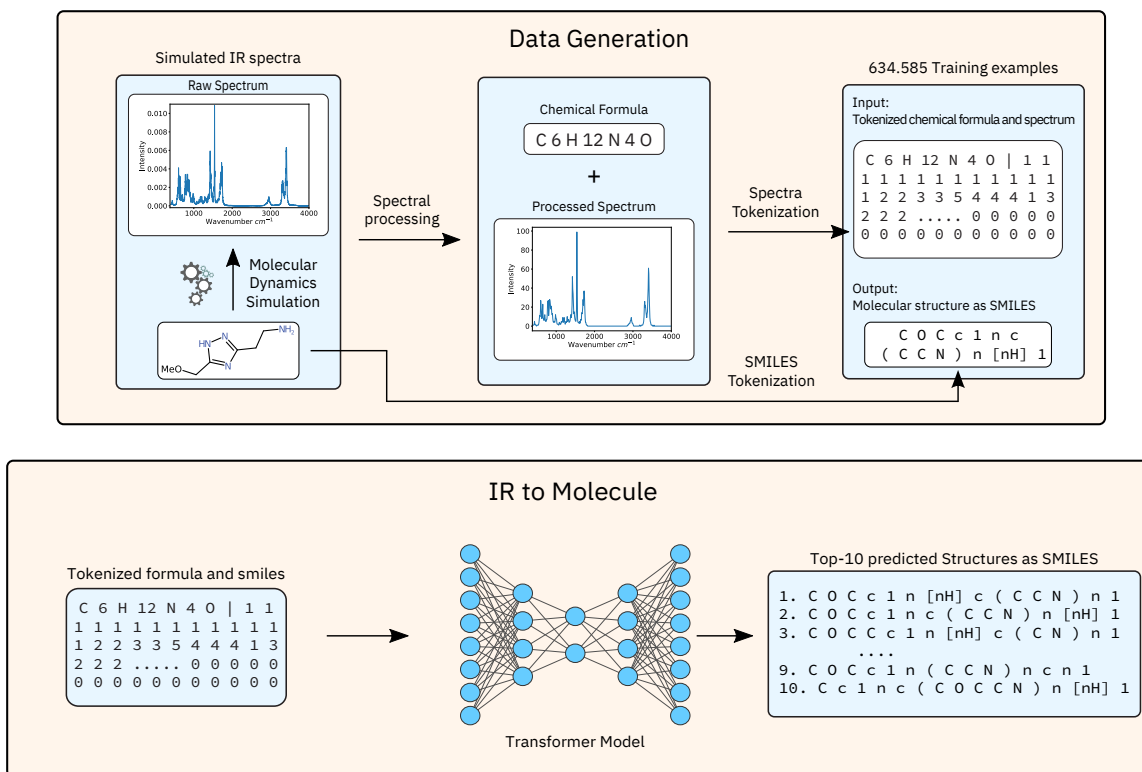


Figure 1: Summary of the processing and prediction pipeline, **Top:** The raw spectrum is processed and together with the molecular formula converted into the input representation. **Bottom:** The input representation is fed into the transformer model to predict molecular structure as SMILES.

### 2.2.1. Chemical Formula

In order to constrain the chemical space explored, we provide both the chemical formula and the IR spectrum as input to the model. An ablation study was conducted to assess the effect of this combination, with three models trained: one solely using the IR spectrum, another relying exclusively on the chemical formula, and a third combining both modalities (see Table 1, “Chemical Formula”). The model incorporating both the spectrum and formula as input outperforms both other models. Conversely, the model trained solely on the chemical formula performed the worst being only able to predict the correct structure in the top-10 in 0.056% of cases. This demonstrates that the model does not only provide reasonable isomers for a given chemical formula but instead is able to learn structural features from the IR spectrum.

### 2.2.2. Sequence length

In order to fine-tune the model, we evaluated the influence of modifying the sequence length used to represent the IR spectrum. The resolution of the spectrum and the degree of information available to the model are both determined by the sequence length. As the length of the sequence increases, transformer models generally exhibit a decrease in

Table 1: Summary of the experiments and associated metrics.

	Formula	Spectrum	Tokens*	Window	Top-1%	Top-5%	Top-10%
Chemical Formula	✓	✗	400	N/A	0.01	0.03	0.06
	✗	✓	400	Full	17.01	33.6	39.37
	✓	✓	400	Full	<b>26.21</b>	<b>51.38</b>	<b>59.63</b>
Token lengths	✓	✓	100	Full	19.17	41.21	49.63
	✓	✓	200	Full	23.14	46.27	54.59
	✓	✓	300	Full	25.78	50.47	58.81
	✓	✓	400	Full	<b>26.21</b>	<b>51.38</b>	<b>59.63</b>
	✓	✓	500	Full	18.03	39.67	47.26
	✓	✓	600	Full	3.36	10.49	14.18
	✓	✓	700	Full	2.31	7.54	10.40
	✓	✓	800	Full	2.22	7.13	9.63
	✓	✓	900	Full	2.13	6.71	9.01
	✓	✓	1000	Full	2.89	9.05	12.40
Windows	✓	✓	400	UM IR <sup>‡</sup>	10.14	26.37	33.98
	✓	✓	400	Fp <sup>†</sup>	25.83	48.39	56.36
	✓	✓	400	Full	26.21	51.38	59.63
	✓	✓	400	Merged <sup>§</sup>	<b>30.22</b>	<b>55.87</b>	<b>63.88</b>
Best Augmented	✓	✓	400	Merged <sup>§</sup>	<b>36.43</b>	<b>62.49</b>	<b>70.00</b>
Best Ensemble	✓	✓	400	Merged <sup>§</sup>	<b>45.33</b>	<b>72.21</b>	<b>78.50</b>

\* Number of tokens encoding the IR spectrum

<sup>†</sup> Fp: Fingerprint, 400–2000 cm<sup>-1</sup>

<sup>‡</sup> UM IR: Upper middle IR, 2000–3982 cm<sup>-1</sup>

<sup>§</sup> Merged: 400–2000 cm<sup>-1</sup> and 2800–3300 cm<sup>-1</sup>

performance, resulting in a broad range of outcomes [27]. Therefore, we studied the effect of varying the number of tokens encoding the spectra from 100 to 1000. This is equivalent to altering the resolution of the spectrum from  $\sim 36$  cm<sup>-1</sup> to  $\sim 3.6$  cm<sup>-1</sup> (see Figure 2 and Table 1, “Token lengths”).

The accuracy slowly increases up to a maximum, followed by a sharp decrease. At low sequence lengths, the input data does not provide sufficient information to allow the model to make an informed decision. On the other hand, the model struggles with longer sequences, as with an increase in resolution a lot of the values in the spectra become redundant and the model has to differentiate between relevant and redundant information. Based on these findings, a sequence length of 400 was chosen for all further experiments.

### 2.2.3. Window selection

Equally important as the sequence length is which part of the spectrum the model is trained on as some section of the IR spectrum, e.g. the fingerprint region contain vastly more peaks than others. In all previous studies, the model was trained on the full spectrum. Here we vary the window, i.e. the specific part of the spectrum which is provided as input. We selected four distinct sets: the full spectrum (resolution:  $\sim 9$  cm<sup>-1</sup>), the fingerprint region (400–2000cm<sup>-1</sup>, resolution: 4 cm<sup>-1</sup>), the upper middle IR region (2000–3982

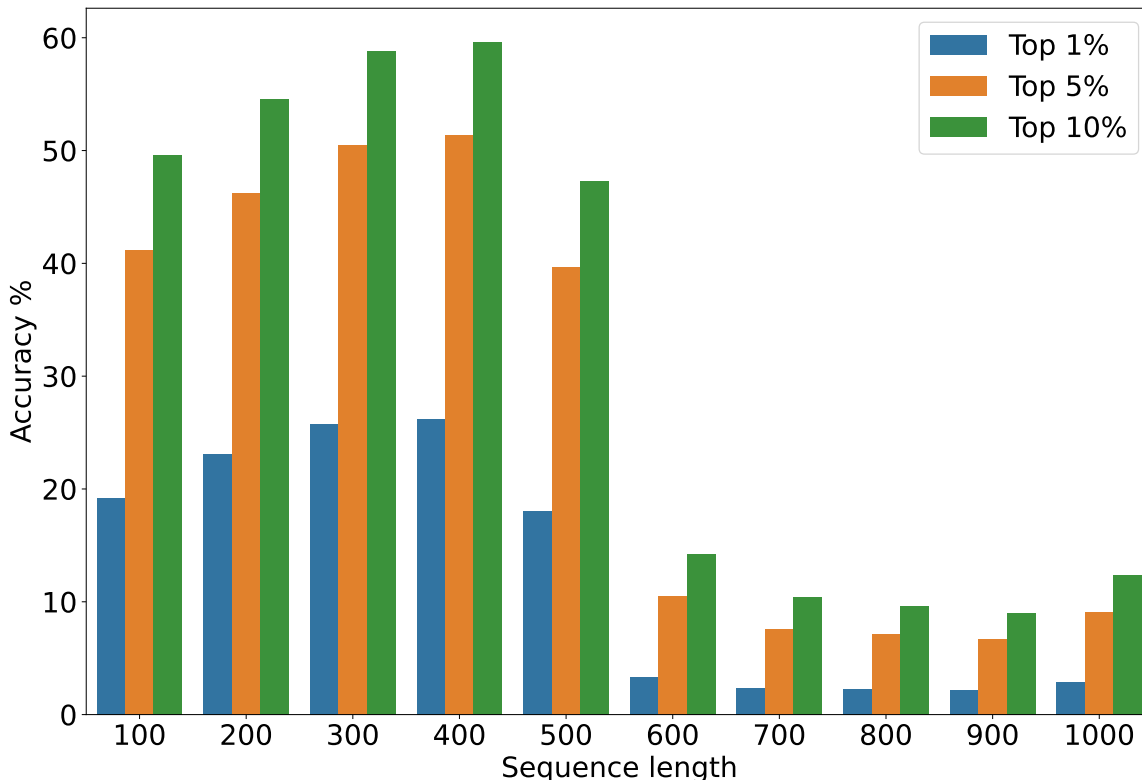


Figure 2: Model accuracy plotted against length of the sequence encoding the IR spectrum.

$\text{cm}^{-1}$ , resolution:  $\sim 5 \text{ cm}^{-1}$ ), and a merged split containing the fingerprint region and a window in the range of  $2800\text{--}3300 \text{ cm}^{-1}$  (resolution:  $5.25 \text{ cm}^{-1}$ ). The number of tokens describing the spectrum was kept constant at 400, causing the resolution to differ from set to set.

After evaluating the performance of each window (see Table 1, “Windows”), we found that the merged split performs best. On the other hand, the upper middle IR region performs the worst as this region mostly consists of hydrogen stretches and overtones. The fingerprint and the full spectrum perform similarly, indicating that the higher detail of the fingerprint region compensates for the loss of information with regards to the full spectrum. This demonstrates that the model is capable of learning the relationship between the anharmonicities of the IR fingerprint region and how they change with the molecular structure. The merged split performs best, likely because it contains both the fingerprint region and the region around  $3000 \text{ cm}^{-1}$  while providing more detail than the full spectrum option.

#### 2.2.4. Data augmentation

The training data was augmented using three methods, described in Methods 4.4: Horizontal shift, vertical noise and smoothing. Table 2 shows the effect of each method separately and the increase in performance of a model trained on data augmented using

all three methods. Each augmentation method increases the data by a factor of two.

Table 2: Results of different augmentation techniques.

	Top-1%	Top-5%	Top-10%
Vertical Noise	6.97	18.53	23.55
No Augmentation	30.22	55.87	63.88
Horizontal shift	<b>36.43</b>	<b>62.49</b>	<b>70.0</b>
Smoothing	35.47	61.1	68.54
Horizontal shift + Smoothing	31.82	58.11	66.17
Horizontal shift + Smoothing (augmented test set)	31.77	58.12	66.13

Adding vertical noise showed a surprisingly strong effect on the model’s performance, significantly degrading it. Our interpretation for this observation is that the model makes use of precise patterns in the shape of the peaks to predict the molecular structure, and that the addition of noise disrupts these patterns, leading to the observed degradation.

In contrast, both horizontal shifting and smoothing the spectrum resulted in a 5-6% increase in performance, with horizontal shifting showing a slight advantage over smoothing. Based on these results, a model was trained using both horizontally shifted and smoothed data. We observed that its performance was comparable to the non-augmented model. These findings prompted us to evaluate the performance of the model on augmented data, which was four times the size of the original data, to ascertain whether the model had become more adept at interpreting the augmented data. However, the evaluation on the augmented data yielded results that were similar to the non-augmented test set. Accordingly, we believe that the decreased performance results from the increased complexity found in the augmented data.

### 2.2.5. Ensembles

To further increase the performance of the model we used an ensemble of the five best performing checkpoints resulting from one training run, and by ensembling two models that contain the weight average of 10 checkpoints of two independently initialized training runs. The best augmented model was utilised (horizontal shift). This further increases the accuracy by 5% and 9% respectively (see Table 3).

Table 3: Results of different ensembling techniques.

	Top-1%	Top-5%	Top-10%
Augmented Model	36.43	62.49	70.0
Ensemble of 5	41.42	68.79	75.46
Ensemble of 2 avg. of 10	<b>45.33</b>	<b>72.21</b>	<b>78.50</b>

## 2.3. Model analysis

In the following, we present an analysis of the model predictions with respect to different characteristics, such as the heavy atom count or the presence of specific functional groups. The results are based on the best-performing model, i.e. the one resulting from ensembling two averaged runs.

### 2.3.1. Heavy atom dependency

To analyze the model’s performance, we evaluated its accuracy against the heavy atom count. Figure 3 shows a negative correlation between the heavy atom count (i.e. all atoms without hydrogen of a molecule) and accuracy. This correlation likely stems from three factors. Firstly, as the heavy atom count increases, molecules become more complex, resulting in longer SMILES strings. Since the model predicts molecules autoregressively, even a single incorrectly predicted token can produce a widely different structure. Secondly, with an increase in the number of atoms, the number of vibrational modes increases according to  $3N - 6$ , where  $N$  is the total number of atoms in the molecule. As a result, the spectrum becomes more crowded, and peaks begin to overlap, which reduces the model’s ability to make accurate predictions. The last factor is that with an increase in the heavy atom count the chemical space increases exponentially. As such, there are more potential isomers that the model has to differentiate, making the prediction more challenging. Figure 9 (see Appendix A) shows the heavy atom count distribution in the test set and reflects this exponential increase. All three factors can be addressed by training the model on more data. More data would also allow for a larger model architecture further increasing the performance.

### 2.3.2. Functional group to structure

Another factor affecting the model’s performance are certain functional groups. We evaluated the model’s ability to predict the correct molecular structure based on the presence of a set of 21 functional groups in the target molecule (see Figure 4). The functional group definitions and results in tabular form can be found in Tables 5 and 6 in Appendices B and C, respectively. To avoid bias caused by the size of the underlying compounds, we calculated the average heavy atom count for molecules containing each of the particular groups. The average heavy atom count for all functional groups falls within  $11.5 \pm 0.7$ .

The model performs best when benzene is present in the target structure, likely due to benzene’s identifiable aromatic C-C stretches. Additionally, the presence of benzene determines a significant portion of the overall structure of the molecule, simplifying the structure prediction task. The good performance of the model on molecules containing cyano and alkyne groups could similarly be explained by these reducing the complexity of the molecule.



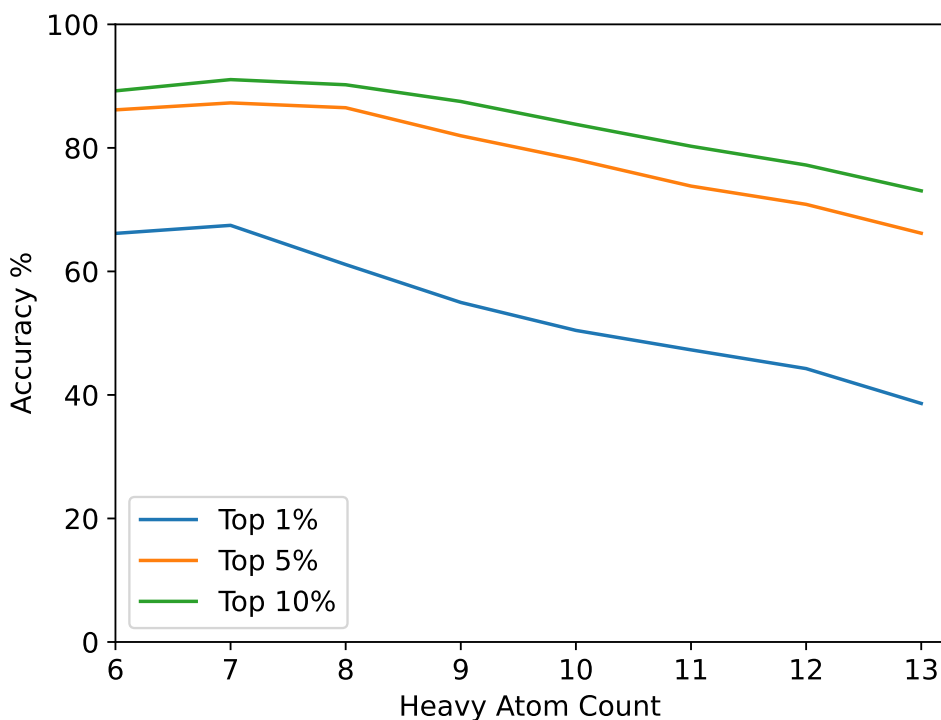


Figure 3: Heavy atom count vs accuracy.

On the other hand, the model performs poorly when predicting the structure of phosphoesters and phosphoric acids. This is likely not an intrinsic limitation of these functional group but the outcome of a low representation of these functional groups in the training data set. Phosphoesters and phosphoric acids are both only present in  $\sim 0.2\%$  of molecules in the training set, while all other functional groups are found in at least  $3\%$  of molecules. This distribution reflects the occurrence of these functional groups in PubChem.

### 2.3.3. Functional group prediction

Previous research has focused on predicting functional groups from the IR spectrum [16–18]. To compare our work in this common task we assessed the model’s performance by comparing the functional groups present in the target molecule with those in the top-1 prediction (see Figure 5). All halogens were excluded from this analysis as their presence in the chemical formula makes the prediction trivial. The model demonstrates high accuracy in predicting the presence of most functional groups, with F1 scores above 0.92 for all functional groups except aldehydes (see Table 7 in Appendix D). Overall, the model has an average F1 score of 0.961 and an average weighted F1 score of 0.970 on the functional groups analysed. The model’s poor performance on aldehydes can be attributed to the fact that the aldehydes are confused with ketones (see section 2.3.4).

It is interesting to observe that while our model was trained on molecular structure prediction, it achieves a high accuracy while predicting functional groups, even outperforming models trained solely to perform this task. Jung et al. achieved a weighted F1

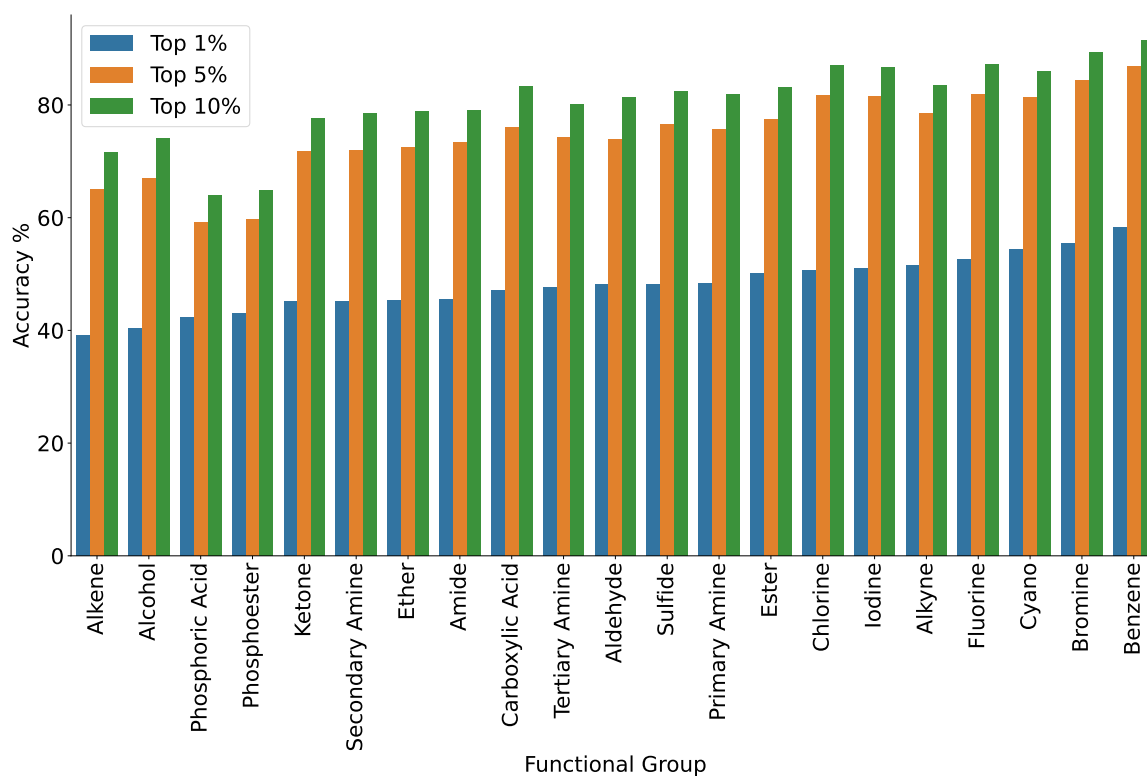


Figure 4: Model accuracy plotted against the occurrence of specific functional group in the target molecule.

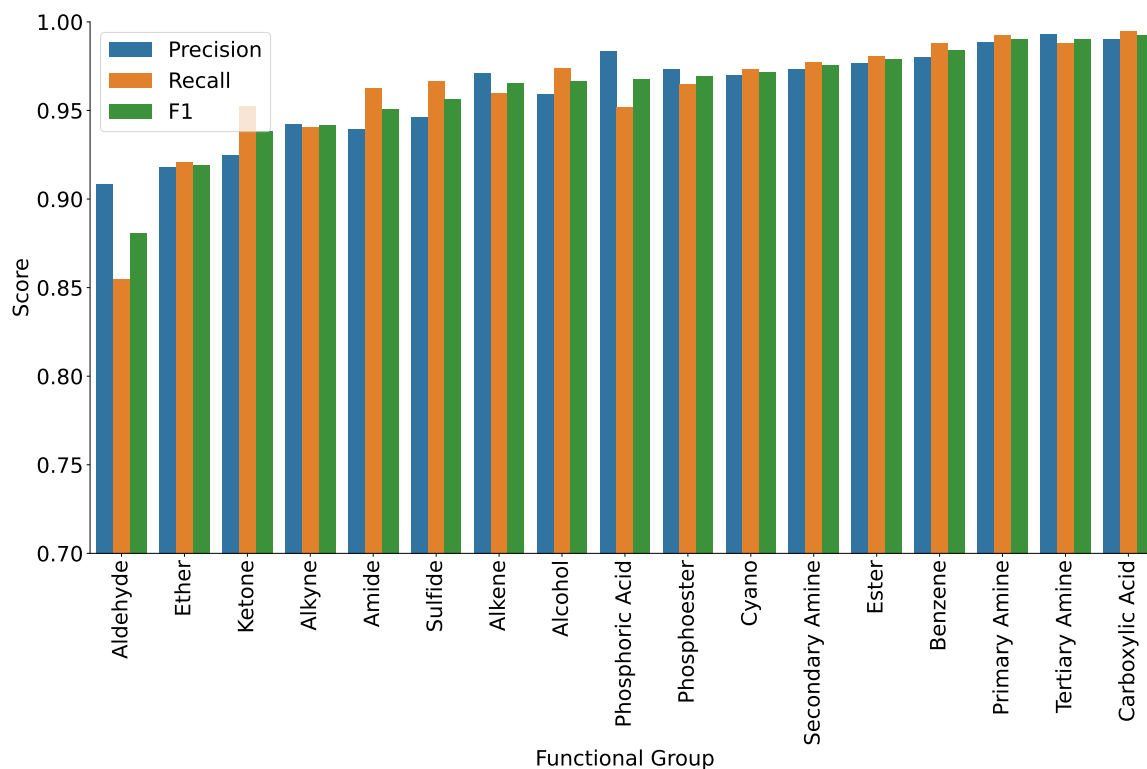


Figure 5: The models ability to predict the functional group in the target group. Calculated by comparing the functional groups of the top-1 prediction to the ground truth.

score of 0.930 using CNNs [17] while Fine et al. demonstrated an average F1 score of 0.926 using MLPs [18]. To keep the results consistent we adopted the exact functional group definitions used in each paper for the following comparison. Our model outperforms both of these previous results (see Table 4). However, it has to be noted that our model includes the chemical formula as additional input and that these results were obtained on a test set with a different functional group distribution.

Functional group definition	Model	Avg. F1	Avg. Weighted F1
Jung (37 groups) [17]	Jung [17]	0.850	0.930
	Ours	<b>0.911</b>	<b>0.976</b>
Fine (15 groups) [18]	Fine [18]	0.926	N/A.
	Ours	<b>0.956</b>	<b>0.978</b>

Table 4: Comparison of our model’s ability to predict functional groups in the top-1 prediction to Jung and Fine’s models solely trained to predict functional groups based on the IR spectrum.

#### 2.3.4. Functional group heat map

To assess why the model makes wrong predictions, we analyse the correlation between the expected functional group and the incorrectly predicted functional group. For this we analyse the set of predictions where the expected functional group is not present in the top-1 prediction (see Figure 6). We calculate the representation of functional groups in the set of false positives for a given functional group compared to the normal distribution of the test set (see Figure 10 in Appendix E). In the heat map, if a functional has a value of six, it represents that it occurs six times more frequently in the set of target molecules where a wrong prediction was made compared to the whole test set. Conversely, a value below zero represents an underrepresentation and zero that the same distribution is found in both the set of wrong predictions and the whole test set.

The heatmap in Figure 6 shows high values where there is a strong correlation between the expected and predicted functional group, i.e. when the model confuses certain functional groups. Interestingly, the model’s confusion patterns align with what one might expect from a human interpreting an IR spectrum. For example, the model often confuses carboxylic acids with esters, which is expected given the similarity of their peaks in the  $1700\text{ cm}^{-1}$  region. Similarly, the model often confuses aldehydes and ketones with each other. The lacking performance of the model on predicting aldehydes, seen in section 2.3.3, can be explained by the fact that the model is twice as likely to confuse aldehydes with ketones as ketones with aldehydes. This confusion stems from two factors: Firstly the the peaks of aldehydes and ketones are very similar and secondly the ketones are two times more common in the training data than aldehydes.

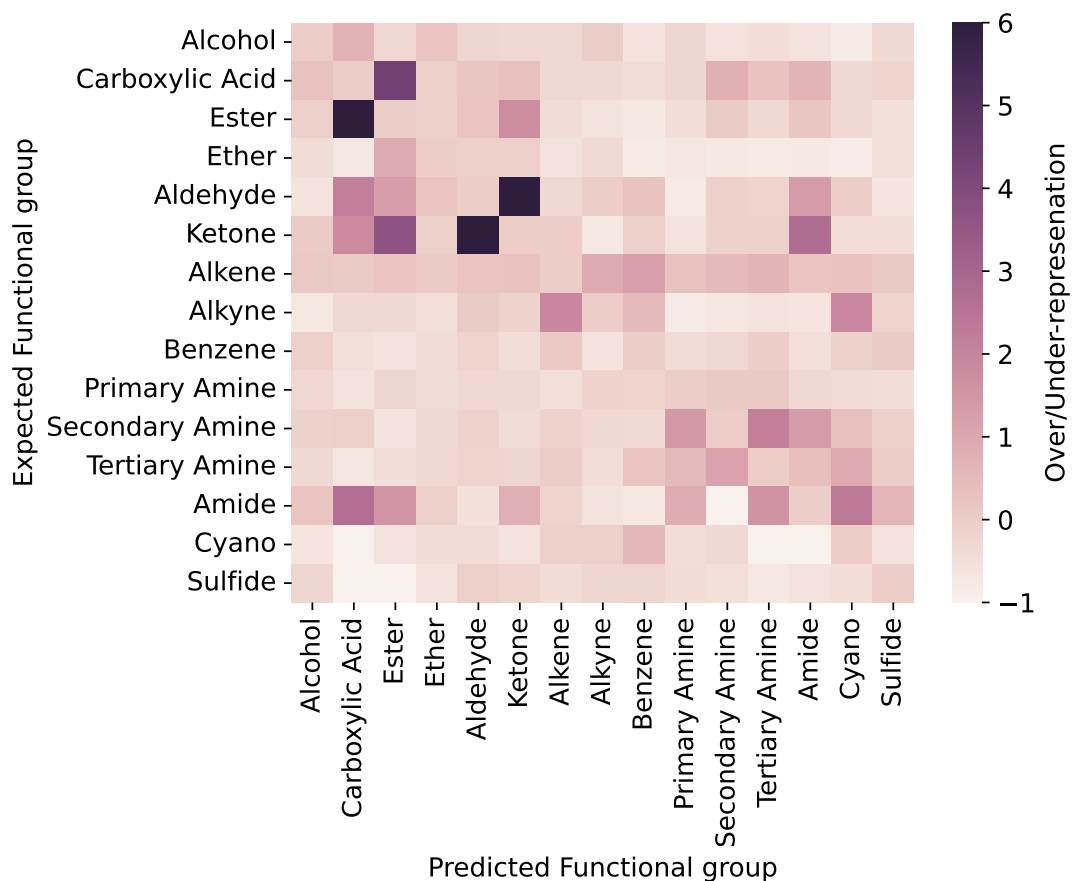


Figure 6: Model accuracy against the occurrence of specific functional group in the target molecule.

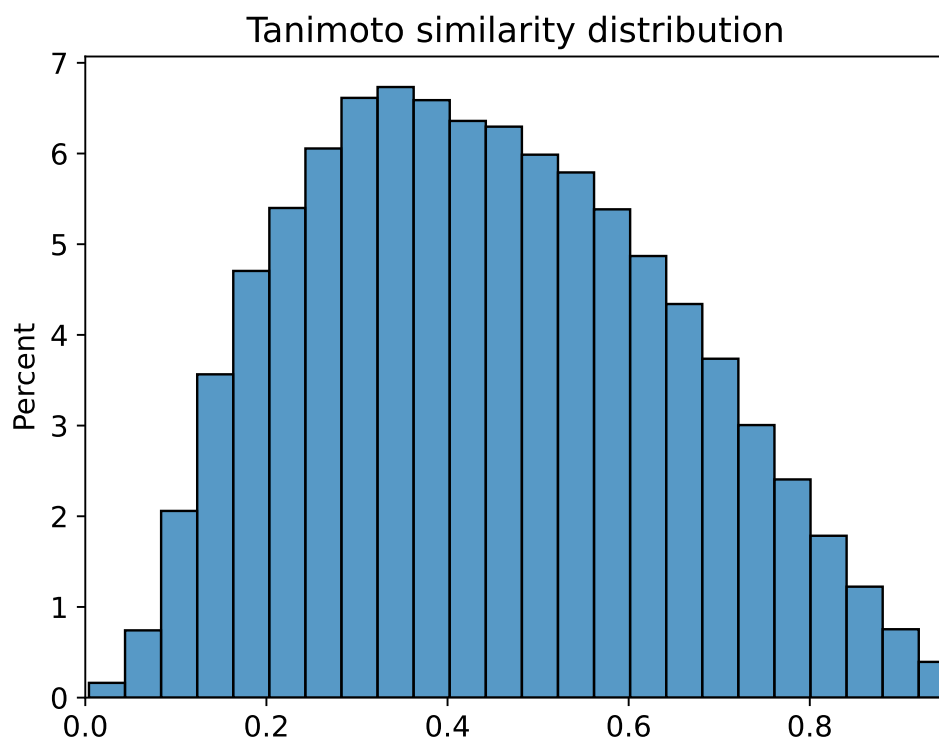


Figure 7: Tanimoto similarity distribution of all top-10 excluding correct ones of the model to the ground truth.

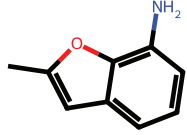
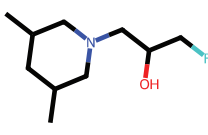
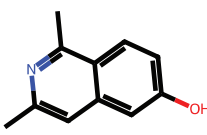
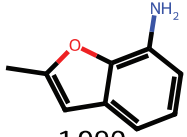
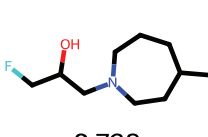
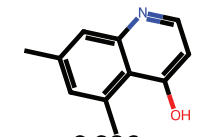
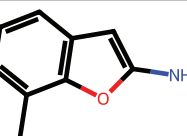
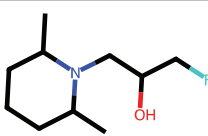
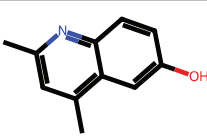
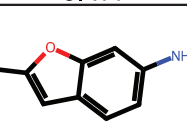
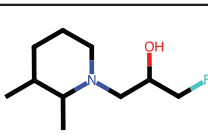
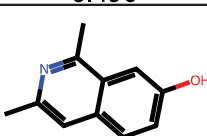
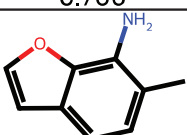
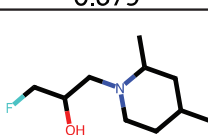
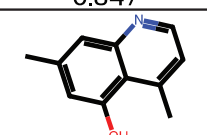
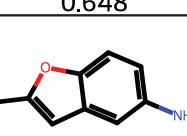
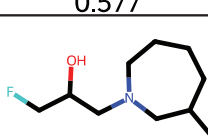
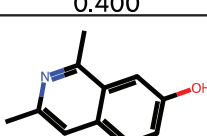
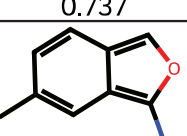
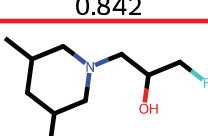
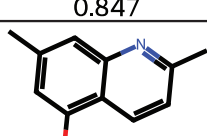
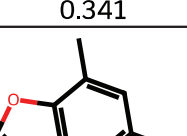
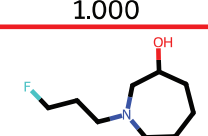
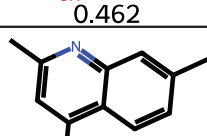
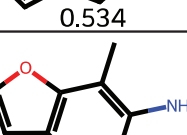
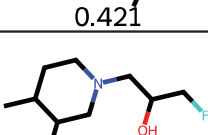
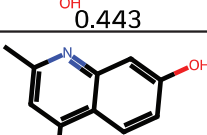
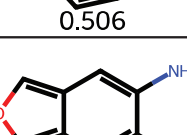
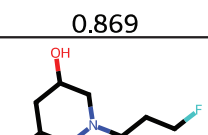
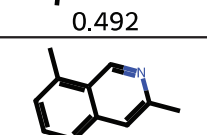
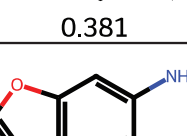
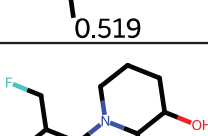
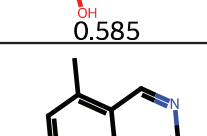
Target			
Pred: 1	 1.000	 0.738	 0.386
Pred: 2	 0.477	 0.547	 0.490
Pred: 3	 0.706	 0.679	 0.847
Pred: 4	 0.648	 0.577	 0.400
Pred: 5	 0.737	 0.842	 0.847
Pred: 6	 0.341	 1.000	 0.462
Pred: 7	 0.534	 0.421	 0.443
Pred: 8	 0.506	 0.869	 0.492
Pred: 9	 0.381	 0.519	 0.585
Pred: 10	 0.532	 0.451	 0.705

Figure 8: Three different examples of the model predicting a molecule from an IR spectrum. The top row consists of the target, with the predictions in ranked order as predicted by the model below. Correct predictions are highlighted in red and the Tanimoto similarity is given below each predicted structure.

### 2.3.5. Similarity

The Tanimoto similarity [28] to the ground truth was calculated for all predicted molecules, excluding correct predictions. A histogram of the similarities is depicted in Figure 7. The mean similarity for the whole set is 0.473. This value represents a relatively high similarity, demonstrating the model’s general understanding of IR spectra.

Also speaking to the model’s ability to predict the correct structure is that the interval between 0 and 0.2 solely accounts for 10.5% of predictions, and as such the vast majority of predicted molecules are similar to the target molecule. Figure 8 demonstrates this for three randomly chosen molecules with all ten predictions showing high similarity to the target.

We also assessed the model’s ability to predict the correct scaffold from the IR spectrum. Murcko scaffold’s were constructed using RDKit [29]. Our model is able to correctly match the scaffold as the top-1 prediction in 77.01%, in the top-5 in 88.79% and in the top-10 in 91.54% of cases.

### 2.3.6. Limitations

One of the key limitations of our methodological approach lies in the availability of datasets containing infrared spectra. While such datasets do exist, licenses for their use are often expensive and restrict machine learning applications, limiting their use. Consequently, we were compelled to simulate IR spectra using molecular dynamics and force fields. While this approach is not inherently limiting, it is important to note that the resulting spectra are highly coherent and consistent. In reality, the situation may be different, and the spectra may exhibit greater variability and inconsistencies. Hence, it is crucial to keep this limitation in mind while interpreting the results of our study.

## 3. Conclusions

We have presented a transformer model that for the first time is capable of elucidating the molecular structure directly from IR spectra. We trained and evaluated our model on a set of simulated IR spectra, sampled using molecular dynamics. Our best model shows a top-1 performance of 45.33%, top-5 of 72.21% and top-10 performance of 78.50%.

We found that our model is able to correctly predict functional groups with an F1 score of 0.961, outperforming all previous works. Upon analyzing the model’s failures in predicting functional groups, we observed that it made errors resembling those a human analyst might commit when interpreting a spectrum. This demonstrates the model’s capability to learn how to interpret spectra.

When making errors, the molecules predicted by the model are very similar to the target molecule with an average Tanimoto similarity of 0.47. In addition, when only considering

the scaffold our model is capable of predicting the correct scaffold in 77.01% and 91.54% of the top-1 and top-10 predictions, respectively.

While our model is solely trained on simulated data, fine-tuning the model on experimental data will allow the model to learn the variability of experimental data while leveraging fundamentals learned from simulated data. We envision a democratization of the structural characterization, a future where a first analysis of an unknown substance could be carried out using IR spectroscopy, instead of requiring expensive and time consuming NMRs. This is especially applicable for research institutions who cannot afford complex and expensive analytical instruments.

## Code availability

The code for generating the data and training the models is available at <https://github.com/rxn4chemistry/rxn-ir-to-structure>.

## Data availability

The IR spectra generated for this work and on which the models were trained are openly accessible at <https://doi.org/10.5281/zenodo.7928396>.

## Acknowledgments

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. We thank Oliver Schilter, Federico Zipoli, Jannis Born, Nicolas Deutschmann, and Amol Thakkar for helpful discussions.

## 4. Methods

### 4.1. Model

We base our model architecture on the Molecular Transformer [13]. The model takes the formatted IR spectrum with the chemical formula as input and outputs a molecular structure encoded as SMILES. This can be formulated as a translation task from the spectrum to the molecular structure. The model is implemented using the standard transformer of OpenNMT-py library [30,31] with the following hyperparameters deviating:

```
word_vec_size: 512  
hidden_size: 512
```

layers: 4  
batch\_size: 4096

All models are trained for 250k steps amounting to approximately 35h on a V100 GPU.

## 4.2. Synthetic data

Using Molecular Dynamics, the spectra of 634,585 molecules were generated. The molecules were sampled from PubChem [32] and filtered to exclude charged molecules, stereoisomers, and molecules containing atoms other than C, H, N, O, S, P, and the halogens, while restricting the heavy atom count to a range of 6–13. A molecular dynamics simulation was run for 800,000 molecules sampled from this set.

A high throughput pipeline was developed in Python to orchestrate molecular dynamics simulations and calculate the spectra from the molecule’s dipole moment. The pipeline utilised EMC [33] to generate the input files for a LAMMPS simulation [34,35]. PCFF is utilised as force field [22]. The system is allowed to equilibrate for 250 ns, before recording the dipole moment of the molecule for a further 250 ns. IR spectra are calculated from the dipole moment according to Braun [36].

A total of 634,585 spectra with a resolution of  $2\text{ cm}^{-1}$  and range of  $400\text{--}3982\text{ cm}^{-1}$  were successfully generated representing a success rate of 75.6%. Most errors were caused by bond types not being parameterised by PCFF.

## 4.3. Data processing

The input of the model consists of the IR spectrum and the molecular formula. The molecular formula is calculated using RDKit. The input representation of the IR spectrum was obtained by interpolating over the specified range and to a given resolution. All spectra are normalized to the range 0–99 and converted to integers. The molecular formula is split into atom types and numbers and the IR spectrum is appended to this string following the vertical bar, “|”, as a separating token. All SMILES strings were canonicalised ensuring a consistent molecular representation and tokenized according to Schwaller et al. [13].

The dataset was split into a train, test and validation set. Test and validation set sizes were chosen as 10% and 5% respectively of the full dataset.

## 4.4. Data augmentation

Both experimental and synthetic data were augmented using Gaussian smoothing, horizontal shifts and vertical noise.

Smoothing was performed using a 1D-Gaussian filter with sigma of 0.75 and 1.25 before interpolating to the desired resolution and window.



As the resolution of the processed input spectrum is in all cases above  $4\text{ cm}^{-1}$ , two shifted spectra can be constructed. The first is the spectrum starting at  $400\text{ cm}^{-1}$  and taking every second value, constructing a spectrum with resolution  $4\text{ cm}^{-1}$ . Similarly, this can be done by taking every second value starting at  $402\text{ cm}^{-1}$ . The input spectra are obtained by interpolating with the desired range and resolution over these two spectra.

To add noise to the spectra, the maximum value of each value was calculated, multiplied by 0.05. For each datapoint in the spectrum this value is further multiplied with a number sampled from a normal distribution with mean 0.00 and standard deviation 0.25 and added to the spectrum. Following this the spectrum was interpolated to afford the desired range and resolution.

Each augmentation technique yields two augmented spectra.

## References

- [1] R. Bowling Barnes and Lyman G. Bonner. The Early History and the Methods of Infrared Spectroscopy. *American Journal of Physics*, 4:181–189, 1936.
- [2] John Coates. Interpretation of Infrared Spectra, A Practical Approach. In *Encyclopedia of Analytical Chemistry*, pages 10815–10837. John Wiley & Sons Ltd, 2020.
- [3] Barbara Stuart. Infrared Spectroscopy. In *Analytical Techniques in Forensic Science*, pages 145–160. John Wiley & Sons, Ltd, 2021.
- [4] Chi-Shi Chen, Yue Li, and Chris W Brown. Searching a mid-infrared spectral library of solids and liquids with spectra of mixtures. *Vibrational Spectroscopy*, 14:9–17, 1997.
- [5] Frank Platte and H. Michael Heise. Substance identification based on transmission THz spectra using library search. *Journal of Molecular Structure*, 1073:3–9, 2014.
- [6] K. Varmuza, P. N. Penchev, and H. Scsibrany. Large and frequently occurring substructures in organic compounds obtained by library search of infrared spectra. *Vibrational Spectroscopy*, 19:407–412, 1999.
- [7] Matthew Gundlach, Katherine Paulsen, Michael Garry, and Steve Lowry. Yin and yang in chemistry education: the complementary nature of FTIR and NMR spectroscopies. Technical report, 2015.
- [8] Andre J. Simpson, Myrna J. Simpson, and Ronald Soong. Nuclear Magnetic Resonance Spectroscopy and Its Key Role in Environmental Research. *Environmental Science & Technology*, 46:11488–11496, 2012.

- [9] Christoph Seger. Usage and limitations of liquid chromatography-tandem mass spectrometry (LC-MS/MS) in clinical routine laboratories. *Wiener Medizinische Wochenschrift*, 162:499–504, 2012.
- [10] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112:12580–12585, 2015.
- [11] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015. arXiv:1512.03385.
- [13] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5:1572–1583, 2019.
- [14] Michael Moret, Irene Pachon Angona, Leandro Cotos, Shen Yan, Kenneth Atz, Cyrill Brunner, Martin Baumgartner, Francesca Grisoni, and Gisbert Schneider. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications*, 14(1):114, 2023.
- [15] Alain C. Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H. Nair, Anna Iuliano, and Teodoro Laino. Inferring experimental procedures from text-based representations of chemical reactions. *Nature Communications*, 12(1):2573, 2021.
- [16] Abigail A. Enders, Nicole M. North, Chase M. Fensore, Juan Velez-Alvarez, and Heather C. Allen. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Analytical Chemistry*, 93:9711–9718, 2021.
- [17] Guwon Jung, Son Gyo Jung, and Jacqueline M. Cole. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical Science*, 14:3600–3609, 2023.
- [18] Jonathan A. Fine, Anand A. Rajasekar, Krupal P. Jethava, and Gaurav Chopra. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical Science*, 11:4618–4630, 2020.
- [19] Kevin Judge, Chris W. Brown, and Lutz Hamel. Sensitivity of Infrared Spectra to Chemical Functional Groups. *Analytical Chemistry*, 80:4186–4192, 2008.

- [20] Christoph Klawun and Charles L. Wilkins. Optimization of Functional Group Prediction from Infrared Spectra Using Neural Networks. *Journal of Chemical Information and Computer Sciences*, 36:69–81, 1996.
- [21] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.
- [22] Huai Sun, Stephen J. Mumby, Jon R. Maple, and Arnold T. Hagler. An ab Initio CFF93 All-Atom Force Field for Polycarbonates. *Journal of the American Chemical Society*, 116:2978–2987, 1994.
- [23] Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling*, 62:2064–2076, 2022.
- [24] Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery, 2019. arXiv:1911.04738.
- [25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929.
- [26] Jannis Born and Matteo Manica. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5:432–444, 2023.
- [27] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer, 2020. arXiv:2004.05150.
- [28] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:20, 2015.
- [29] RDKit. <https://www.rdkit.org/> (Accessed April 14, 2023).
- [30] OpenNMT-py: Open-Source Neural Machine Translation, 2017. <https://github.com/OpenNMT/OpenNMT-py> (Accessed April 20, 2023).
- [31] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation, 2017. arXiv:1701.02810.

- [32] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51:D1373–D1380, 2023.
- [33] Pieter J. in 't Veld and Gregory C. Rutledge. Temperature-Dependent Elasticity of a Semicrystalline Interphase Composed of Freely Rotating Chains. *Macromolecules*, 36:7358–7365, 2003.
- [34] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022.
- [35] LAMMPS Molecular Dynamics Simulator. <https://www.lammps.org> (Accessed April 20, 2023).
- [36] Efrem Braun. Calculating An IR Spectra From A Lammeps Simulation, 2016. 10.5281/ZENODO.154672.
- [37] Daylight: SMARTS Examples. [https://www.daylight.com/dayhtml\\_tutorials/languages/smarts/smarts\\_examples.html](https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html) (Accessed April 20, 2023).

## Appendix

### A. Heavy atom count distribution

The heavy atom count of the test set is shown in Figure 9. As the molecules were randomly sampled from PubChem [32], this distribution reflects the distribution of the PubChem database. The number of molecules increases exponentially with the heavy atom count, demonstrating the exponentially larger chemical space with an increase in the heavy atom count.

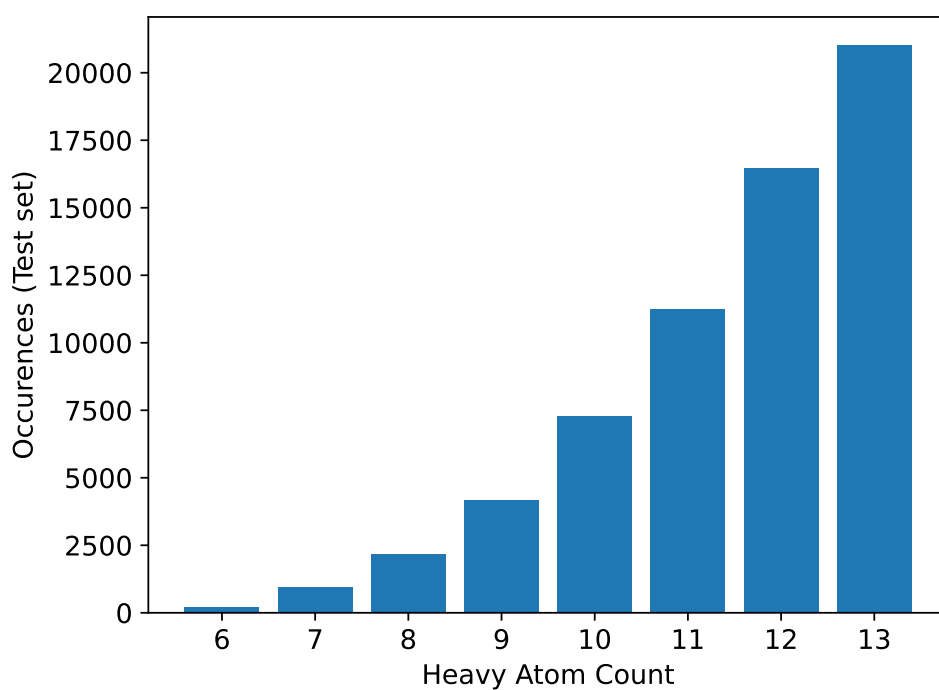


Figure 9: Heavy atom distribution of the test set

## B. Functional group definitions

Functional groups are defined in SMARTS as shown in Table 5. Using these SMARTS and RDKit the presence of a certain function group is determined by invoking `<RDKit molecule>.GetSubstrucMatches(<RDKit molecule from SMARTS patter>)`

Table 5: Functional group definitions used.

	Definition
Alcohol	<chem>[OX2H][CX4;!\$(C([OX2H])[O,S,#7,#15])]</chem>
Carboxylic Acid	<chem>[CX3](=O)[OX2H1]</chem>
Ester	<chem>[#6][CX3](=O)[OX2H0][#6]</chem>
Ether	<chem>[OD2]([#6])[#6]</chem>
Aldehyde	<chem>[CX3H1](=O)[#6]</chem>
Ketone	<chem>[#6][CX3](=O)[#6]</chem>
Alkene	<chem>[CX3]=[CX3]</chem>
Alkyne	<chem>[\$([CX2]#C)]</chem>
Benzene	<chem>c1ccccc1</chem>
Primary Amine	<chem>[NX3;H2;!\$(NC=[!#6]);!\$(NC#[!#6])][#6]</chem>
Secondary Amine	<chem>[NH1,nH1]</chem>
Tertiary Amine	<chem>[NH0,nH0]</chem>
Amide	<chem>[NX3][CX3](=[OX1])[#6]</chem>
Cyano	<chem>[NX1]#[CX2]</chem>
Fluorine	<chem>[#6][F]</chem>
Chlorine	<chem>[#6][Cl]</chem>
Iodine	<chem>[#6][I]</chem>
Bromine	<chem>[#6][Br]</chem>
Sulfonamide	<chem>[#16X4]([NX3])(=[OX1])(=[OX1])[#6]</chem>
Sulfone	<chem>[#16X4](=[OX1])(=[OX1])([#6])[#6]</chem>
Sulfide	<chem>[#16X2H0]</chem>
Phosphoric Acid <sup>†</sup>	<chem>[\$(P(=[OX1])([\$([OX2H]),\$([OX1-]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2]P))])]</chem>
Phosphoester <sup>†</sup>	<chem>[\$(P(=[OX1])([OX2][#6])([\$([OX2H]),\$([OX1-]),\$([OX2][#6])]))([\$([OX2H]),\$([OX1-]),\$([OX2][#6]),\$([OX2]P)))([\$([OX2H]),\$([OX1-]),\$([OX2][#6])]))([\$([OX2H]),\$([OX1-]),\$([OX2][#6])])][\$([OX2H]),\$([OX1-]),\$([OX2][#6]),\$([OX2]P)))]</chem>

<sup>†</sup> Adapted from [37]

### C. Results: Model accuracy depending on specific functional groups

In Table 6 the accuracy of the model depending on the presence of specific functional groups in the target molecule is shown. Count represents the number of molecules with this functional group in the test set. Additionally, the average heavy atom count (Avg. HAC in the table) is calculated to rule out bias.

Table 6: The model’s ability to predict the correct molecular structure based on if a specific functional group is present in the target molecule.

	Count	Avg. HAC	Top-1%	Top-5%	Top-10%
Alkene	17,581	11.18	39.18	64.99	71.60
Alcohol	9,987	11.63	40.31	66.95	73.99
Phosphoric Acid	125	11.48	42.40	59.20	64.00
Phosphoester	114	11.53	42.98	59.65	64.91
Ketone	5,429	11.61	45.15	71.69	77.62
Secondary Amine	19,129	11.59	45.24	72.00	78.51
Ether	10,671	11.86	45.40	72.45	78.93
Amide	1,261	12.14	45.60	73.35	79.06
Carboxylic Acid	4,323	11.87	47.17	76.04	83.25
Tertiary Amine	28,807	11.64	47.64	74.30	80.14
Aldehyde	2,449	11.40	48.18	73.91	81.30
Sulfide	5,723	11.46	48.26	76.57	82.49
Primary Amine	11,848	11.61	48.30	75.67	81.92
Ester	3,409	11.98	50.13	77.44	83.19
Chlorine	3,330	10.71	50.72	81.71	86.99
Iodine	1,748	11.27	51.03	81.46	86.67
Alkyne	2,302	11.32	51.56	78.50	83.49
Fluorine	5,974	11.66	52.58	81.80	87.21
Cyano	3,508	11.58	54.45	81.33	85.97
Bromine	2,880	11.67	55.38	84.38	89.27
Benzene	9,436	12.35	58.34	86.91	91.39

## D. Results: Functional group accuracy

In Table 7 we present the models ability to predict functional groups based on the IR spectrum and the chemical formula. To calculate the scores we compare the functional groups present in the target molecule to those present in the top-1 prediction. Precision, recall and the F1-score were calculated.

Table 7: The model’s ability to predict that a functional group is present based on the IR spectrum. This table is based on the models top-1 prediction.

	Count	Precision	Recall	F1
Aldehyde	2,431	0.91	0.85	0.88
Ether	10,647	0.92	0.92	0.92
Ketone	5,406	0.92	0.95	0.94
Alkyne	2,287	0.94	0.94	0.94
Amide	1,255	0.94	0.96	0.95
Sulfide	5,670	0.95	0.97	0.96
Alkene	17,456	0.97	0.96	0.97
Alcohol	9,961	0.96	0.97	0.97
Phosphoric Acid	125	0.98	0.95	0.97
Phosphoester	114	0.97	0.96	0.97
Cyano	3,475	0.97	0.97	0.97
Secondary Amine	18,991	0.97	0.98	0.98
Ester	3,405	0.98	0.98	0.98
Benzene	9,343	0.98	0.99	0.98
Primary Amine	11,799	0.99	0.99	0.99
Tertiary Amine	28,451	0.99	0.99	0.99
Carboxylic Acid	4,301	0.99	0.99	0.99



## E. Correlation of functional groups in the test set

Figure 10 shows the correlation between individual groups in the test set.

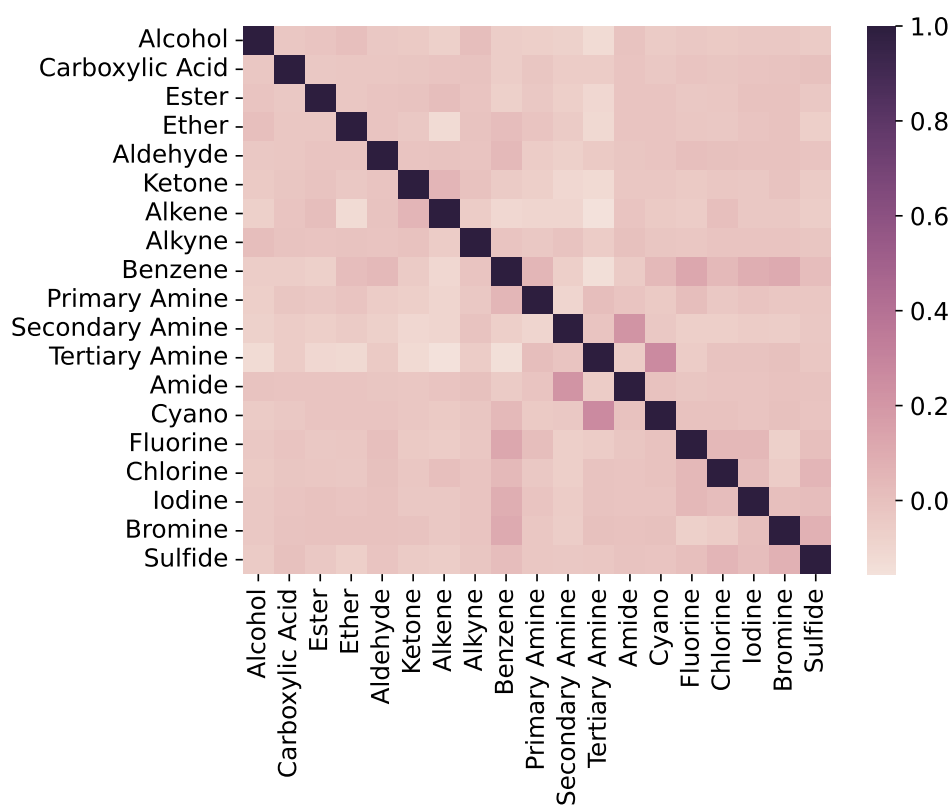


Figure 10: Correlation between functional groups of the test set